Categorization by Reference: A Novel Approach to MeSH Term Assignment

Vram Kouramajian, Ph.D.[†*]
Vijayanth Devadhar[†]
[†]Department of Computer Science
Wichita State University
Wichita, Kansas 67260
{vram,vsdevadh,sxmaram}@cs.twsu.edu

Jerry Fowler, Ph.D.[‡]
Srinivas Maram[†]
[‡]Department of Community Medicine
Baylor College of Medicine
Houston, Texas 77030
gfowler@bcm.tmc.edu

Categorization by Reference is a novel text classification technique that examines the existing classifications of the citations found in an as-yet unclassified text to determine what terms should be assigned to that text. The existence of the Medical Subject Headings and MEDLINE make the biomedical domain a prime candidate for application of this technique. We describe our approach and implementation of a prototype, presenting some results of our initial tests. We further discuss refinements that could improve the precision of the technique, and describe its possible use in categorizing portions of the World-Wide Web.

INTRODUCTION

In biomedicine, as in any scientific domain, published results are the backbone of further research. Naturally, efficient storage and retrieval of their content is of great interest to academics and clinicians. The National Library of Medicine (NLM) employs many full-time domain experts who examine most new academic publications in the field of biomedicine, manually assigning Medical Subject Headings (MeSH terms) to each. Optimizing the time of these experts is essential, given the immense volume of new research being generated and the increasing necessity of timely access to the latest research.

Automation of the tasks of indexing and search is becoming vital to the effort of keeping abreast of current research. As larger fractions of the research corpus become available electronically, new techniques for automating the text categorization process become feasible. Furthermore, although human expertise is ultimately essential to the task of MeSH indexing, distinguished experts in information storage and retrieval assert that human effort alone may not give the best results [1]. Evidence suggests that automated categorization techniques now rival the work of human experts, and can at least provide valuable expert assistance to human indexers [2].

We are developing a new technique for automatic categorization of text based on the citations of related work made by the text itself. To assign keywords to a target document, we analyze the keywords that have previously been assigned to documents that are cited in the target document. The new method, called *Categorization by Reference*, is generally applicable in all fields of intellectual endeavor, but thanks to the efforts of the NLM and the American Medical Informatics Association (AMIA), the field of medical informatics is rife with opportunities for research in automated text categorization. Consequently, our experiments with Categorization by Reference take advantage of the MeSH terms assigned in MEDLINE.

This paper describes Categorization by Reference and presents preliminary experimental results obtained using the method. First, we define the problem more precisely as a question of identifying semantic locality within the research corpus. Second, we describe our method and present some results obtained by selecting an arbitrary set of citations from the proceedings of the 1993 Symposium on Computer Applications in Medical Care (SCAMC) [3]. We then examine issues raised by use of the method, discussing limitations and possible enhancements. Our conclusion describes our belief in the increasing feasibility of the method in the future.

SEMANTIC LOCALITY

Grouping objects by semantic locality (that is, categorizing or differentiating them) is a fundamental intellectual activity addressed by the indexing structure of the Unified Medical Language System (UMLS) Metathesaurus [4]. Categorization by Reference is a method of extracting knowledge about conceptual locality (synonymy, lexical variance, and foreign-language equivalence). It can be extended to deal with contextual locality (relationship within contexts such as the MeSH hierarchy) and occurrence locality (the co-

*Current Affiliation: Knight-Ridder Information, Inc., 2440 El Camino Real, Mountain View, CA 94040.

occurrence of MeSH terms in MEDLINE citations) as well.

"One judges a person by the company he keeps." Categorization by Reference, in analogy with this adage, judges a research paper by the references it cites in its bibliography (the paper's *reference set*). This idea establishes a new measure of conceptual semantic locality that we call *subject intensiveness*:

> The predominant topics of a scholarly work are likely to be the same as those of the papers in the work's reference set.

For example, by perusing the reference set for *this* paper, one can glean from the references to Salton [1], Yang [2], and Schwartz [5] that we discuss methods for information storage and retrieval, the general subject of all those works.

The text in a research paper is normally devoted to a central concept. The idea grows in depth rather than in breadth. The research normally concentrates on extending from a small domain. The reference set will also deal in the immediate proximity of the central concept, because an author will generally not refer to papers grossly unrelated to the work at hand. The subject intensiveness assumption suggests that looking at the reference set allows one to infer the topic of the citing document.

We exploit the semantic locality of subject intensiveness by computing a set of MeSH terms derived from the papers in the reference set. Since the reference set of a new paper in biomedicine must necessarily precede it temporally, the papers in the reference set will typically already have been assigned MeSH terms, which are retrievable automatically from MEDLINE.

Categorization by Reference is a departure from previous work, because we avoid content analysis, so the method is domain *independent*. The learning in Categorization by Reference is implicit in the use of previously assigned MeSH terms. Early methods of mapping from natural language vocabulary to subject categories have typically employed syntactic approaches alone [6]. Current work deals with this problem by adding semantic criteria to classification technique. Most solutions use some sort of knowledge–based system [7, 8]. Knowledge–based approaches suffer from the difficulty of developing machine learning, and also tend to be domain dependent.

Classifying a paper based on its reference set may provide insights that semantic analysis of the text might obscure. For instance, many papers in medical informatics employ the vocabulary of medicine in order to provide concrete examples. For example, a previous publication of ours uses several terms from the field of gynecology. The title and topic of the paper is, in fact, "Modeling Past, Current, and Future Time in Medical Databases." The medical vocabulary used in that paper is relevant to the topic of the paper only for illustrative purposes. This is reflected in that fact that the paper contains no references to research in gynecology (we do not include the citation of that paper in this one because it is mentioned here *only* as an illustration, and bears no other relevance).

CATEGORIZATION BY REFERENCE

In an abstract sense, we can view the collection of documents represented by a set of research papers as a hypertext that has not been computerized. We might think of each reference to another paper as a hypertext link; when a citation is encountered while reading a physical paper copy of the work, the effect of "pressing the mouse button" to activate the link is accomplished by going to the card catalog, library serials locator list, and thence to the library's stacks in search of the cited article. With this abstraction in mind, we find an analogy to clustering methods used to group the points in the characteristic feature space (the "nearest neighbors" technique is commonly used for clustering in pattern recognition [9]). Botafogo has used clustering to improve data display, browsing, and retrieval in hypertext [10].

Algorithm

To assign indexing terms to a given target article, we collect the papers in its reference set, and retrieve all the indexing terms assigned to each of these papers, yielding the *term candidate set* (note that a single term may appear more than once in the term candidate set, once for each member of the reference set in which it appears). We then analyze the terms in the term candidate set using a parameterized term–assignment function to assign scores to each term in the term candidate set. Once the scores have been assigned, those terms scoring highest are assigned as terms to the target article.

The term–assignment function scores individual terms according to several criteria. The primary scoring criterion is frequency of occurrence. A secondary criterion is the time difference between the target article and the reference. In addition, if the knowledgebase that provides the indexing terms differentiates in weight among the terms it returns, then that is also incorporated into the term assignment function. In this experiment, we distinguish between major terms and minor terms as assigned by MEDLINE. Term assignment is governed by the following parameters:

- *Major–Term Acceptance Threshold*, K ($0 \leq K \leq 1$), is the cut-off value above which the term under consideration should be assigned to the target article as a major MeSH term.
- *Minor–Term Acceptance Threshold*, k ($0 \leq k \leq K$), is the minimum value at which a term qualifies as a minor term for the target article. Changing the minor term scoring

factor changes the relative relevance between major and minor terms. Alteration of the acceptance thresholds affects the *recall* and *precision* of retrieval on citations indexed using the algorithm; higher values of K and k increase precision.

- *Age Weighting Factor*, w ($0 \leq w \leq 1$), is the relevance decay factor for a term based on the age of the citation from which it was retrieved. A value of 1 ignores the effect of age, while a value of 0 ignores any but the most recent articles.
- *Minor-Term Weight Factor*, m ($0 \leq m \leq 1$), is the relative importance to be assigned to MeSH terms declared "minor" as compared with those that have been declared "major" for a given reference. A value of 1 equates minor and major terms, while a value of 0 ignores minor terms altogether.
- *Normalized Frequency*, $\mathcal{F}(t,y) = \frac{n}{N}$, where n is the number of times a tuple of the form (t = term, y = year) appears in the reference set of the article, and N is the total number of references in the reference set. Frequency is normalized to overcome the variations in number of documents cited from one target article to another.

The resulting threshold, age factor, weight factor, and normalized frequency are converted to a score $S(t)$ for each term t in the term candidate set using the following function:

$$S(t) = \sum_{y=1}^{n} (F(t,y) + f(t,y) * m) * W(Y_n - y)$$

where $F(t,y)$ is the normalized frequency of use of t as a major term during year y, $f(t,y)$ is the normalized frequency of use of t as minor term during year y, m is the minor term weighting factor, and $W(Y_n - y)$ is the age weighting function. In our experiments, we let $W(age) = w^{age}$ reflect exponential decay of relevance.

When all term candidates have been scored, those scoring more than k are assigned as terms of the target. Terms with scores greater than K are major terms and rest minor terms.

EXPERIMENTAL METHOD

Articles in Medical journals are the ideal candidates for this experiment, since reviewers assign MeSH terms to each document entered in MEDLINE. For our experiment, we selected an arbitrary set of citations from the proceedings of the 1993 SCAMC, which had already been indexed in MEDLINE. This allowed us to compare our computed results with those terms actually assigned.

The testing process comprised four phases: input (scanning), parsing, retrieval of parsed citations, and application of the algorithm.

Scanning. A set of documents called the initial set was selected from the SCAMC proceedings and scanned. This scanned output was human edited for any scanning errors.

Parsing. The initial set was parsed to produce the required format for MEDLINE input. Since there is little adherence to standard in the citation of references in technical papers, our parser uses the following set of heuristics, which accommodated roughly 95% of the references we encountered.

- The format of a reference can be either title first or name first.
- The authors' names can occur either initial first or name first.
- The first blank space not followed by a delimiter in the citation text is assumed to be the beginning of the title.

Citation Retrieval. To get the indices of the cited documents we queried a local copy of MEDLINE using the title and authors parsed from the reference set of each citation, retrieving only the MeSH terms assigned to each member of the reference set. These MeSH terms were grouped in the *term candidates* file.

The term candidates were then processed to assign weights. Varying the thresholds for these term weights, affected the selection of suitable terms to be used as the indices for the citing document.

Results

We evaluated the results of our experiment by comparing MeSH terms assigned by our algorithm with the MeSH terms assigned by human experts (as retrieved through MEDLINE). In our experiments, we varied the values of K, k, m, and w. Here we describe the effect of aging on comprehensiveness of term assignment.

In the experimental runs of Table 1, we fixed the values of K, k, and m, and allowed w to range between 0 (only the most recent references), 0.98 (small decay factor), and 1 (no decay factor). We categorized our experimental input into three classes (I_1, I_2, and I_3) based on the number of citations in sample papers. Papers in class I_1 have few references, in class I_2 have a moderate number of references, and in class I_3 have many references. The two sides of the table reflect the degree to which the algorithm duplicated the work of human MeSH indexers, expressed as percentages. The columns labelled "recall" indicate the fraction of human-assigned terms found in the computed set. The "precision" columns indicate the fraction of computed terms that matched the human-assigned term set. A value of $K = 0.1$ appears to extract essentially all major MeSH-terms (labeled "Major") and a small fraction of minor MeSH-terms (labeled "minor") at the expense of low discrimination. The low overall percentages for minor MeSH-terms could be explained by the

Input	"Recall"						"Precision"		
	w=0		*w=0.98*		*w=1*		*w=0*	*w=0.98*	*w=1*
	Major	minor	Major	minor	Major	minor	Major	Major	Major
I_1	100	0	100	100	100	100	13	21	25
I_2	66	0	100	0	100	0	14	22	22
I_3	28	50	42	50	56	50	3	12	6

Table 1: Percent correlation of computed terms with assigned terms ($K = 0.1, k = 0.05, m = 0.5$).

fact that many minor terms (relating to sponsoring agencies and the like) are irrelevant to semantic locality. The age decay factor is signficant, especially among papers with a large number of references. Clearly, older references are essential to categorization by reference, but "precision" may be improved by decreasing the importance given to older citations.

We also noticed that for papers having few references, our results were closer to assignments made by human experts. This we attribute to our conjecture that, in general, if an author cites only a few references, they tend to be more subject intensive; whereas, if an author cites a large number of papers, these papers tend to be less related. It may be appropriate to choose different values of K for reference sets of different sizes. In our refinement section, we discuss a number of methods that may substantially improve these results.

DISCUSSION

The technique of Categorization by Reference is already practiced by thousands of users of internet newsgroup readers in evaluating which "threads" of discussion to follow. Because the subject lines of articles posted to the newsgroups frequently cite previous articles as references, users can choose to ignore the current article and follow other more interesting threads. Users of some news readers have the ability to eliminate from consideration every submission by a specified author: This is an extreme form of Categorization by Reference.

A similar idea to Categorization by Reference was pursued by Schwartz to characterize a group of people with similar interests or expertise [5]. Samples of electronic mail were collected from 15 sites around the Western Hemisphere and the "To/From" logs were analyzed in order to compute the *interest distance* between individuals.

Our automated tool for Categorization by Reference relies on the ability to parse citations embedded in the text of an article. This requires that citations be provided in a consistent form. Future authors' adherence to citation standards for MEDLINE in the case of biomedical works and to the Modern Language Association style guide for other works will improve the utility of this work. In the absence of standards, we have been forced to use heuristics in parsing citations.

Minor MeSH terms are an overloaded concept, used both for description of the article's environment (by means of terms such as "Support U S Govt P H S") and for themes of minor importance in the paper itself. For this reason, we are developing a stop–list of MeSH terms that should not be assigned automatically no matter how highly they may score.

Categorization by Reference is a descriptive system only. The use of this system cannot expand the vocabulary of the indexing system. Intelligent intervention is still needed to determine when a new category has been developed. However, we conjecture that the inability of Categorization by Reference to assign a definitive set of terms to an article might indicate that the article introduces a new subject area, for which a new index term might be appropriate (this assumes that a low scoring term assignment was *not* caused by inadequate citation of prior work on the part of the article's author, a judgment requiring intelligence).

Refinements

Numerous refinements of this method are possible. We discuss a few here:

The term–assignment function requires a number of parameters. To permit experimentation with different values, the system could provide improved feedback that would help in choosing better parameters. It may even be possible to automate the parameter generation based on criteria such as the average number of terms associated with citations. The use of dynamic thresholds instead of fixed values of K and k may provide improved "precision" [13].

Our current implementation simply reads the "References" section of the target article to find citations for the reference set. This does not distinguish among citations based on where they were footnoted in the text of an article. It may be worthwhile to examine whether the location of a reference has any bearing on the weighting that its index terms should receive. We speculate that references footnoted in the introduction are more general, and perhaps less significant as indicators of the subject of the target paper, than those cited in the body of the article. Also, an article that is mentioned more than once should probably be weighted more heavily than one that is mentioned only once. Bernstein examined the use of contextual information to affect the weight of terms based on their location in a text [11].

Categorization by Reference uses only first–

generation references; that is, MeSH terms derived directly from the reference set. It may be interesting to examine whether the use of more generations (the reference sets of the members of the reference set) affects the quality of the categorization. We hope to compare term–assignment that uses only first–generation references with term–assignment that includes second–generation references also; we believe that, in general, these two sets of MeSH terms will share most of their terms (a large discrepancy might indicate the development of a new topic in biomedicine).

Closely related terms like "Information Systems," "Clinical Information Systems," and "Hospital Information Systems" are often assigned to different articles in the reference set. Since it is generally inappropriate to assign more than one of these terms, and also desirable to assign the most specific applicable term, these relationships should be resolved by making use of the MeSH hierarchy to determine similarity, based on the distance of terms from a common ancestor. Brute–force comparison of all terms for relationships would be prohibitively expensive. We have not yet decided on an appropriate solution.

Categorization by Reference can be combined with other classification techniques such as Latent Semantic Indexing (LSI) [12] or ExpNet [2]. This technique is also applicable to analysis of documents published on the World–Wide Web; although there is no formal attribute assignment for Web pages as there is for MEDLINE, one can still draw conclusions about pages based on interest, for instance, creating a list of "interesting" pages and classifying as "probably also interesting" any page that refers to one or more of these "interesting" pages. In conjunction with the Web–MeSH Medibot, which seeks to assign MeSH terms to Web pages using a term–mapping knowledgebase extracted from MEDLINE [13], Categorization by Reference could assist in identifying clusters of pages related to a specific MeSH term or term set.

SUMMARY

Categorization by Reference is a simple but effective technique relying on the natural clustering of a text with the articles its authors have declared relevant by their citation of them.

Although the technique of categorization by reference is promising, it is not yet universally feasible, even in the medical informatics corpus, because of the difficulty of acquiring electronic versions of research publications. This condition is changing rapidly however: For instance, a requirement of submission to the "MEDINFO'95" conference was that the submission be accompanied by an electronic version of the text, so that the proceedings of the conference could be made available on a Compact Disk Read–Only Memory (CD–ROM). We hope to use this CD–ROM to demonstrate the utility of our method on a larger scale.

We believe that, with refinement, Categorization by Reference will develop into a useful automated aid for the important work of text categorization in medical and other scientific literature.

REFERENCES

[1] Salton G. Another Look at Automatic Text Retrieval Systems. In *Communications of the ACM*, volume 29, pages 648–656, July 1986.

[2] Yang Y., Chute C. An Application of Expert Network to Clinical Classification and MEDLINE Indexing. In *Proc Annu Symp Comput Appl Med Care*, pages 157–162, November 1994.

[3] Editor: Safran, C. Proc annu symp comput appl med care, November 1993.

[4] Nelson S. J., Tuttle M. S., Cole W. G., et al. From Meaning to Term: Semantic Locality in the UMLS Metathasarus. In *Proc Annu Symp Comput Appl Med Care*, pages 209–213, 1991.

[5] Schwartz M. Internet Resource Discovery at the University of Colorado. In *IEEE Computer*, volume 26, pages 25–35, September 1993.

[6] Maron M. Automatic Indexing: An Experimental Inquiry. In *Journal of the ACM*, volume 8, pages 404–417, 1961.

[7] Apte C., Damerau F., Weiss S. M. Automated Learning of Decision Rules for Text Categorization. In *ACM Trans Information Systems*, volume 12, pages 233–251, July 1994.

[8] Liddy E. D., Paik W., Yu E. S. Text Categorization for Multiple Users Based on Semantic Features from a Machine-Readable Dictionary. In *ACM Trans Information Systems*, volume 12, pages 278–295, July 1994.

[9] Fu K. S. *Syntactic Pattern Recognition and Applications*. Prentice-Hall Inc., 1982.

[10] Botafogo R. A. Cluster Analysis for Hypertext Systems. In *ACM SIGIR Conf Research and Development in Information Retrieval*, pages 116–124, June 1993.

[11] Bernstein L. M., Williamson R. E. Testing of a natural language retrieval systems for a full text knowledge base. In *Journal American Soc Information Science*, volume 35, pages 235–247, 1984.

[12] Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R. Indexing by Latent Semantic Analysis. In *Journal American Soc Information Science*, volume 41, pages 391–407, 1990.

[13] Fowler J., Kouramajian V., Maram S., Devadhar V. Automated MeSH indexing of the world–wide web. In *Proc Annu Symp Comput Appl Med Care*, Oct. 1995. In press.